DOCUMENT RESUME

ED 473 565 TM 034 741

AUTHOR	van der Linden, Wim J.; Veldkamp, Bernard P.
TITLE	Constraining Item Exposure in Computerized Adaptive Testing with Shadow Tests. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
SPONS AGENCY	Law School Admission Council, Princeton, NJ.
REPORT NO	RR-02-06
PUB DATE	2002-00-00
NOTE	34p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; College Entrance Examinations; *Computer Assisted Testing; Law Schools; Probability; Test Construction; *Test Items
IDENTIFIERS	*Item Exposure (Tests); *Law School Admission Test

ABSTRACT

Item-exposure control in computerized adaptive testing is implemented by imposing item-ineligibility constraints on the assembly process of the shadow tests. The method resembles J. Sympson and R. Hetter's (1985) method of item-exposure control in that the decisions to impose the constraints are probabilistic. However, the method does not require time consuming simulation studies to set values for control parameters prior to the operational use of the test. Instead, the probabilities of item ineligibility can be set "on the fly" using an adaptive procedure based on the actual item-exposure rates. An empirical study using an item pool from the Law School Admission Test showed that application of the method yielded perfect control of the item exposure rates and had negligible impact on the bias and mean squared error functions of the ability estimates. (Contains 5 figures and 26 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM TMR 0077

ED 473 565

Constraining Item Exposure in Computerized Adaptive Testing with Shadow Tests

Research Report 02-06

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Nelissen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Wim J. van der Linden
Bernard P. Veldkamp

TM034741

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

BEST COPY AVAILABLE

Constraining Item Exposure in Computerized Adaptive Testing with Shadow Tests

Wim J. van der Linden
Bernard P. Veldkamp

This study received funding from the Law School Admissions Council (LSAC). The opinions and conclusions contained in this paper are those of the other and do not necessarily reflect the policy and position of LSAC. The authors are most indebted to Wim M. M. Tielen for his computational assistance. Requests for reprints should be sent to W.J. van der Linden, Department of Educational Measurements and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, THE NETHERLANDS. Email: w.j.vanderlingen@edte.utwente.nl

Abstract

Item-exposure control in computerized adaptive testing is implemented by imposing item-ineligibility constraints on the assembly process of the shadow tests. The method resembles Sympson and Hetter's (1985) method of item-exposure control in that the decisions to impose the constraints are probabilistic. However, the method does not require time-consuming simulation studies to set values for control parameters prior to the operational use of the test. Instead, the probabilities of item ineligibility can be set "on the fly" using an adaptive procedure based on the actual item-exposure rates. An empirical study using an item pool from the Law School Admission Test (LSAT) showed that application of the method yielded perfect control of the item exposure rates and had negligible impact on the bias and MSE functions of the ability estimator.

Key words: computerized adaptive testing; item-selection constraints; item-exposure control; randomized control; shadow test approach.

Constraining Item Exposure in
Computerized Adaptive Testing with Shadow Tests

Computerized adaptive testing (CAT) can be viewed as an item-selection process in which an objective function is optimized subject to several constraints. A popular objective function in CAT is Fisher's information measure at the estimated ability level of the examinee (for alternative functions see, van der Linden & Pashley, 2000). Constraints have to be imposed on the item-selection process to give the test the necessary composition with respect to such attributes as item content, format, gender orientation, word counts, and statistical item parameters. It is the purpose of this paper to show that the set of constraints can also be extended with constraints that control the exposure rates of the items in the test.

The optimization problem involved in CAT can be represented by a test assembly model with 0-1 decision variables for the items. Let $i = 1, ..., I$ denote the items in the pool. Decision variable x_i is defined to take the value 1 if this item is selected in the test and the value 0 if item i remains in the pool. Suppose that the objective is to maximize Fisher's information measure, which takes the value $I_i(\theta)$ for item i at ability level θ. In addition, suppose that the constraints have to control the composition of the tests with respect to a set of categorical attributes (e.g., item content and format) and quantitative attributes (e.g., statistical item parameters and word counts). The categorical attributes partition the item pool into subsets of items with the same value on the attribute, whereas the quantitative attributes have numerical values. We will use V_g, $g = 1, ..., G$, as a generic symbol to denote subsets of items that share a categorical attribute and q_i as a generic symbol to denote the value of item i on a quantitative attribute.

A general representation of the optimization problem in CAT is:

$$\text{maximize} \sum_{i=1}^{I} I_i(\theta) x_i \qquad (1)$$

subject to

$$\sum_{i \in V_g} x_i \leq u_g, \quad g = 1, ..., G,$$

$$\sum_{i \in V_g} x_i \geq l_g, \quad g = 1, ..., G,$$

$$\sum_{i=1}^{I} q_i x_i \leq u_q,$$

$$\sum_{i=1}^{I} q_i x_i \geq l_q,$$

$$\sum_{i=1}^{I} x_i = n,$$

$$x_i \in \{0, 1\}, \quad i = 1, ..., I. \tag{2}$$

The objective function maximizes the test information at the examinee's true ability value. The first two sets of constraints require the number of items from the set V_g to be between upper bound u_g and lower bound l_g. Likewise, the next two constraints require the sum of values of attribute q_i to be between upper and lower bound u_q and l_q. The final two constraints define the length of the test and the range of the decision variables. For examples of more specific types of constraints that may be needed in test assembly models for CAT, see van der Linden (1998; 2000).

Shadow Test Approach

If all items in the test were to be selected simultaneously and θ could be set to a known value, the model in (1)-(2) could be solved immediately for an optimal set of values for

the decision variables x_i, $i = 1, ..., I$, using a general-purpose software packages for 0-1 linear programming (LP), for example, the software package CPLEX 6.6 (ILOG, Inc., 2000). These values tell us which set of items constitutes the optimal test. However, in adaptive testing items are selected one at a time and each next item has to be optimal at an estimate of θ that changes during the test.

An effective way to solve the assembly model for adaptive tests is through a shadow test approach (van der Linden, 2000; van der Linden & Reese, 1998). In this approach, prior to the selection of each item, the model in (1)-(2) is solved for a full-size linear test (=shadow test) at the current ability estimate with the decision variables of the items already administered fixed to the value 1. The CAT algorithm then picks the optimal item at the ability estimate from the free items in the shadow test for administration. Because each shadow test meets the constraints, the adaptive test automatically meets them. Likewise, because each shadow test is optimal at its ability estimate and the optimal item at the estimate is administered, the adaptive test is optimal.

On-line assembly of shadow tests has become possible through recent developments in 0-1 LP. The techniques in this area have become so fast that, for a current PC, it takes less than a second to assemble a test from an item pool of the size typically used in CAT (Veldkamp, 2001). However, the shadow test approach is a general scheme for optimizing the selection of items in CAT; it can also be used in combination with other techniques for optimal test assembly, such as the algorithms proposed by Luecht (1998) and Swanson and Stocking (1993).

Number of Constraints and Feasibility

For a typical test assembly problem, the number of constraints in the model needed to represent all test specifications can easily run into the hundreds. A set of values for the decision variables that meets all constraints is known as a feasible solution. Algorithms in 0-1 LP algorithms search for a set of values that has the best value for the objective function among the feasible solutions. This solution is known as an optimal feasible solution.

In principle, LP algorithms are able to solve problems with unlimited numbers of constraints. However, a possible effect of adding a constraint to the model is a decrease in the value of the objective function for the optimal solution. The tighter the constraints, the larger the effect can be. It is even possible to overconstrain the model, in which case no feasible solution is left and the LP problem is said to be infeasible.

Feasibility of an LP model is a property that depends only on the constraints. In an adaptive test, the only quantity in the model for the shadow test that changes during the test is the objective function. Therefore, if the model for the shadow tests in (1)-(2) is feasible for any examinee, it is feasible for all examinees. Also, feasibility is maintained during the adaptive test because the items for which the decision variables are fixed at $x_i = 1$ were part of a previous solution. In practical applications, feasibility of test assembly problems is never a problem. If it would, the event should be taken to signal a flaw in the composition of the item pool relative to the specifications of the test (van der Linden, Veldkamp, & Reese, 2000; Veldkamp & van der Linden, 2000).

Item-Exposure Control

Item-exposure control in CAT is necessary to prevent possible item compromise due to overexposure of items. Methods for item-exposure control based on Sympson and Hetter's (1985) proposal have become the *de facto* standard of the CAT industry. These methods are based on the definitions of two events: (1) item i is selected by the CAT algorithm (S_i) and (2) item i is administered (A_i). Because

$$A_i \subset S_i, \tag{3}$$

it holds that

$$P(A_i) = P(A_i \mid S_i)P(S_i) \tag{4}$$

The probabilities $P(S_i)$ are determined by the composition of the item pool and the nature of the CAT algorithm. However, the conditional probabilities $P(A_i \mid S_i)$ are control parameters that can be set to guarantee that

$$P(A_i) \leq r_{max}, \; i = 1, ..., I, \tag{5}$$

with r_{max} being a target value determined by the testing agency. In the Sympson-Hetter methods, after an item is selected the control parameters, $P(A_i \mid S_i)$, are implemented through a probability experiment conducted to determine if the item is actual administered. If an item is not administered, it is removed from the pool and the experiment is repeated for the next best item.

Admissible values for the control parameters in the Sympson-Hetter method can not be calculated analytically. Instead, they have to be found through a series of iterative adjustments, with each iteration step being based on a sufficiently large number of simulated administrations of the adaptive test. Though it is possible to modify the Sympson-Hetter method to speed up the iterative adjustment process considerably (van der Linden, 2002), the process is generally tedious and time consuming, particularly if the control parameters have to be set conditional on a set of realistic ability values for the population of examinees.

For empirical studies on the Sympson-Hetter method or other implementations of the idea to determine item administration probabilistically, see Davey and Parshall (1995, April), Eggen (2001), McBride and Martin (1983), Parshall, Hogarty and Kromrey (1999, June), Revuelta and Ponsoda (1998), van der Linden (2002), and van der Linden and Veldkamp (2002).

Constraining Item Exposure Rates

It is the purpose of this paper to present an alternative method of item-exposure control that is not based on the idea of controlling for item exposure after an item is selected but *before* an examinee takes the test. That is, the decisions to be made are no

longer if an item that is selected should be administered, but which of the items in the pool are eligible for the examinee. If an item is eligible it remains in the pool; if it is ineligible it is removed from the pool for the examinee. The idea underlying the method in this paper is to base the decisions on the outcomes of a probabilistic experiment with probabilities of eligibility that constrain the item-exposure rates to be below the target value.

The main advantage of the method is that no time consuming simulation studies are necessary to find admissible values for control parameters of items. Instead, it is possible to implement the method "on the fly" during operational testing. The method automatically adapts the probabilities of item eligibility to their optimal level and maintains these during the rest of the testing process.

Ineligibility Constraints

A natural way to implement control of item eligibility in adaptive testing with shadow tests is through the inclusion of constraints in the assembly model in (1)-(2). If the decision is made that item i is ineligible for the current examinee, the following constraint is added to the model for the shadow tests:

$$x_i = 0. \tag{6}$$

If item i remains eligible, no constraint is added. We will refer to the constraints in (6) as *ineligibility constraints*.

As noted earlier, a potential effect of adding constraints to a test assembly model is infeasibility of the model. In principle, infeasibility may thus occur if a large number of ineligibility constraints is added to the model. However, for a real-life CAT program, the probability of running into infeasibility due to item ineligibility constraints is generally small. The reason is that the method in this paper imposes ineligibility constraints only for the items that have a tendency to be overexposed. It is a common experience in adaptive testing that this number is low. In fact, the majority of the items in real-life CAT programs are hardly exposed at all.

If the addition of ineligibility constraints results in an infeasible model for the shadow tests, a simple measure to restore feasibility is to remove all ineligibility constraints from the model and solve the relaxed model. The amount of time needed to detect infeasibility of the model is negligible. Besides, as will be shown below, the adaptive nature of the probabilistic experiment in this paper results in an automatic correction for an occasional extra exposure of an item due to infeasibility later in the testing process.

Probabilistic Constraints on Item Eligibility

The process involved in selecting a tests for an examinee in adaptive testing with shadow test is as follows: First, for each item in the pool a probability experiment is conducted to determine which items are eligible and which are ineligible; Second, ineligibility constraints are added to the model for the shadow tests and the model is solved for optimal shadow tests at the sequence of updated ability estimates during the test. Three, if the addition of the ineligibility constraints leads to an infeasible model, the constraints are removed from the model and the relaxed models for the shadow tests are solved. Fourth, from each shadow test, the (free) item with maximum information at the ability estimate is administered.

In summary, the process of item administration is thus defined by the following events:

E_i: item i is determined to eligible for the examinee by the probability experiment;

F: the model for the assembly of the shadow test for the examinee remains feasible after the eligibility constraints have been added;

S_i: item i is selected in a shadow test for the examinee;

A_i: item i is administered to the examinee.

The event of removing the ineligibility constraints from the model is equivalent to $\overline{F}$. Observe that S_i is now used to denote selection of the item in a shadow test and not selection for administration.

Analogous to (3)-(4), it now holds that

$$A_i \subset S_i \subset \{E_i \cup \overline{F}\}. \tag{7}$$

and

$$P(A_i) = P(A_i \mid S_i)P(S_i \mid E_i \cup \overline{F})P(E_i \cup \overline{F}). \tag{8}$$

However, because

$$P(A_i \mid S_i)P(S_i \mid E_i \cup \overline{F}) = P(A_i \mid E_i \cup \overline{F}),$$

we may ignore event S_i and (8) reduces to

$$P(A_i) = P(A_i \mid E_i \cup \overline{F})P(E_i \cup \overline{F}). \tag{9}$$

The problem is to select values for the probabilities of eligibility, $P(E_i)$, such that the exposure rate for item i is below r. Combining (5) with (9), it follows that the target r_{max} for the item exposure rate for item i is met if:

$$P(E_i \cup \overline{F}) \leq \frac{r_{max}}{P(A_i \mid E_i \cup \overline{F})}. \tag{10}$$

However, this inequality does not impose any direct constraint on $P(E_i)$.

We therefore make the following independence assumption

$$P(E_i \cap F) = P(E_i)P(F). \tag{11}$$

In addition, from (7)

$$P(A_i \cap (E_i \cup \overline{F})) = P(A_i).$$

Using

$$P(E_i \cup \overline{F}) = 1 - P(\overline{E}_i \cap F), \tag{12}$$

and

$$P(\overline{E}_i) = 1 - P(E_i),$$

it follows that (10) holds if:

$$P(E_i) \leq 1 - \frac{1}{P(F)} + \frac{r_{max} P(E_i \cup \overline{F})}{P(A_i)P(F)}, \; P(A_i) > 0, P(F) > 0. \tag{13}$$

Because we want the exposure rates of the items with a tendency to overexposure to be just below the target in (5), the probabilities of item eligibility should be set at the upper bound in (13).

Observe that these probabilities are all marginal with respect to the distribution of θ. To obtain constraints that impose item exposure control conditional on θ (Stocking & Lewis (1998), all probabilities should be replaced by conditional probabilities given θ. This operation is straight-forward, and its results are not presented here. We will return to the issue of conditional item-exposure control later in this paper.

Adapting the Probabilities of Item Eligibility

The idea is to set this upper bound adaptively during the test. That is, after an examinee is tested the probabilities of the events in the right-hand side of (13) are updated using the events recorded for the examinee. More specifically, assume that j examinees have been tested and $j+1$ is the next examinee. The probabilities are then updated by

$$P^{(j+1)}(E_i) = \min \left\{ 1 - \frac{1}{P^{(j)}(F)} + \frac{r_{max} P^{(j)}(E_i \cup \overline{F})}{P^{(j)}(A_i)P^{(j)}(F)}, 1 \right\}, \tag{14}$$

provided $P^{(j)}(A_i) > 0$ and $P^{(j)}(F) > 0$.

The following argument helps to understand (14). As already noted, for a well-designed real-life CAT program, the probability of infeasibility is small. That is, we

expect the testing procedure to approximate

$$P^{(j)}(F) = 1 \tag{15}$$

and

$$P^{(j)}(E_i \cup \overline{F}) = P^{(j)}(E_i). \tag{16}$$

Under these conditions, it follows from (14) that

$$\begin{aligned} P^{(j+1)}(E_i) &< P^{(j)}(E_i) \quad \text{if} \quad P^{(j)}(A_i) > r_{\max}, \\ P^{(j+1)}(E_i) &= P^{(j)}(E_i) \quad \text{if} \quad P^{(j)}(A_i) = r_{\max}, \\ P^{(j+1)}(E_i) &> P^{(j)}(E_i) \quad \text{if} \quad P^{(j)}(A_i) < r_{\max}. \end{aligned} \tag{17}$$

These relations show the behavior we want the probabilities of item eligibility to have: If the probability of exposure for an item is too high at some stage during the testing process, its probability of eligibility should go down. If the probability of exposure for an item is already below the target, the probability of eligibility should be relaxed. If the conditions in (15)-(16) do not hold, the relation between $P^{(j+1)}(E_i)$ and $P^{(j)}(E_i)$ is slightly more complicated because the probability of feasibility intervenes and the actual probability of item i being available for selection for examinee j is not $P^{(j)}(E_i)$ but $P^{(j)}(E_i \cup \overline{F})$.

The adaptive nature of the update in (14) also explains our earlier claim that an occasional extra exposure of some of the items in the pool due to the addition of ineligibility constraints to the model for the shadow tests is automatically corrected later in the testing process. Using (12), the expression in the right-hand side of (14) can be written as

$$1 - \frac{1}{P^{(j)}(F)} \left(1 - \frac{r_{\max}}{P^{(j)}(A_i)} \right) - \frac{r_{\max} P^{(j)}(\overline{E}_i)}{P^{(j)}(A_i)}. \tag{18}$$

If infeasibility occurs for examinee j, $P^{(j)}(F)$ decreases. The effect of this decrease is an increase in the probabilities of eligibility, $P^{(j+1)}(E_i)$, in (14), provided $P^{(j)}(A_i) < r_{\max}$.

Independence Assumption

This assumption of independence in (11) is the only empirical assumption on which the method is based. The assumption is generally realistic. Unless the item pool is badly designed, it typically has multiple sets of items with the combinations of attributes required by the content constraints in the test assembly model. The probability of a feasible solution therefore does not depend on the ineligibility of a small number of the items in the pool.

The assumption can easily be tested for an operational CAT program by a scatter plot for the items in the pool with estimates of the probabilities $P(E_i \cap F)$ against products of the estimates of $P(E_i)$ and $P(F)$. Deviations from the identity line would point at violations of the assumption of independence . If some of these probabilities are extremely low, we could plot estimates of $\ln P(E_i \cap F)$ against $\ln P(E_i)$ and check for nonlinearity.

Implementing the Method

Suppose that for the previous j examinees we have recorded the following counts:

φ_j: number of examinees for which the shadow test has been feasible (event F);

ρ_{ij}: number of examinees for which item i has been eligible or the ineligibility constraints have been removed after infeasibility (event $E_i \cup \overline{F}$);

α_{ij}: number of examinees to which item i has been administered (event A_i).

Thus, for examinee $j+1$, item i is eligible with estimated probability

$$\widehat{P^{(j+1)}(E_i)} = \min\left\{1 - \frac{j}{\varphi_j} + \frac{j\rho_{ij}r_{\max}}{\alpha_{ij}\varphi_j}, 1\right\}, \tag{19}$$

with $\alpha_{ij} > 0$ and $\varphi_j > 0$. The decision to add an ineligibility constraint for item i to the model for the shadow test is based on a probability experiment with this probability.

Because of the adaptive nature of the method, the method can be applied directly to an operational test. The only conditions that should be avoided for the estimates in (19), is α_{ij} =0 and $\varphi_j = 0$. A simple way to avoid these conditions is to initialize $\widehat{P^{(j+1)}(E_i)}$ at 1 and keep it at this value long as the counts α_{ij} and φ_j are equal to zero. This measure was taken in the empirical examples below.

Stabilization of Probability Estimates

If the method is applied directly to an operational test, the relative counts in (19) stabilize if the number of examinees tested increases. As long as these counts are unstable, some of the estimates of the probabilities of eligibility may fluctuate considerably. This behavior is particularly likely for items that are favored by the CAT algorithm (e.g., items with high values for the discrimination parameter and values for the difficulty parameter near the center of the ability distribution). These items will be the first to become ineligible. Also, as soon as they are eligible again, they tend to be administered. If the relative counts become stable, the opposite phenomenon will be observed. The probability estimates in (19) then hardly change and items will remain eligible or ineligible for long periods.

If a more consistent behavior of the estimates of the probabilities of item eligibility during the testing process is desirable, they should be updated using the technique of fading introduced in the literature on Bayesian networks (Jensen, 2001, sect. 3.3.2). In this technique counts of events are updated by combining new data with old data weighted by a fading factor chosen from the interval (0,1). Let w_i be the choice for the factor for item i. A common factor w for all items will suffice in most applications. If j examinees have been tested, the weighed updates of the count of the events A_i is given by the following relation:

$$\alpha^*_{i(j+1)} = \begin{cases} w_i\alpha^*_{ij} + 1 & \text{if } i \text{ was administered to } j, \\ w_i\alpha^*_{ij} & \text{if } i \text{ was not administered to } j, \end{cases} \tag{20}$$

where the asterisk is used to denote a weighted versions of the counts. The updates of the weighted counts φ^*_j and ρ^*_{ij} are analogous, whereas the weighted count of the number of examinees is updated by

$$j^*_{(j+1)} = w_i j^*_{ij} + 1. \tag{21}$$

If these weighted counts are substitute into (19), the impact of old events fades away during the testing process. In fact, the choice of weight w_i implies an effective sample

size that tends to $1/(1-w_i)$ (Jensen, 2001, sect. 3.3.2). As a consequence, the estimates of the probabilities of item eligibility tend to be based permanently on the last $1/(1-w_i)$ examinees.

If operational testing should begin with stable values for the probabilities in (19), this goal can be realized by simulating the adaptive test on the computer with weighted updates of the counts for $1/(1-w_i)$ examinees prior to the start of the testing process. In the empirical example below, we used weights $w_i = .999$, which amounted to an effective sample size of 1,000 examinees.

Item Sets

In adaptive testing, item pools sometimes have a structure with sets of items organized around common stimuli (e.g., text passages or descriptions of cases). It is possible to apply the techniques of item-exposure control in this paper at the level of individual items within sets. However, in test with item sets the concern typically is about overexposure of stimuli rather than items. It is therefore recommended to apply the techniques in this paper at the level of the stimuli. Because the exposure rates of items in sets can never be larger than the rates of their stimuli, the criterion in (5) is automatically satisfied for the items if it is for the stimuli.

Empirical Examples

A simulation study was conducted to check the efficacy of the method of exposure control presented in this paper and estimate its effects on the statistical properties of the ability estimator. The following conditions were compared:

1. CAT without item-exposure control;

2. CAT with item-exposure control with the probabilities of item eligibility updated without fading;

3. CAT with item-exposure control with the probabilities of item eligibility updated with fading.

The condition without item-exposure control was included to provide a base line for the evaluation of the performances of the exposure control methods in the other conditions. For the two conditions with exposure control, the target for the maximum exposure rate was set at $r = .25$, a choice believed to be typical of actual targets in real-life CAT programs. For the condition with fading, factor w_i in (20)-(21) was set equal to .999.

Item Pool and Test

The item pool was a previous 753-item pool from the Law School Admission Test (LSAT). All items fitted the 3-parameter logistic item response model (Hambleton & Swaminathan, 1985). The pool had both discrete items and items organized as sets around a common stimulus. The exposure of the items in the sets was at the level of the stimuli. The total number of discrete items and stimuli in the pool was 353.

A 50-item adaptive version of the LSAT was simulated. The length of this test was half the length of the paper-and-pencil version of the LSAT; all specifications for the LSAT were therefore scaled back to 50%. The LSAT consists of three different section, where two of the sections have an item-set structure. The total number of constraints in the LP model for the shadow tests needed to represent all test specifications with respect to such attributes as item and stimulus content, (sub)types, gender and minority orientation, answer key, word count, was equal to 433.

CAT Algorithm

Test administrations were simulated for examinees randomly sampled from a uniform distribution over θ=-2.0, -1.5, ...,2.0. For the two conditions with the method in presented in this paper we used 1, 000 replications to study stabilization of the probability updates, and then another 9,000 replications to get the results presented below. Sampling from a uniform distribution of θ was chosen to enable the estimation of the bias and mean-squared error (MSE) functions of the estimator of θ with equal precision along its scale (1,000 replications for each θ value). The items were selected using the maximum-information criterion. The estimator of θ was the expected a posterior (EAP) estimator with an

uninformative prior over [-4,4]. In all conditions, the ability estimate was initialized at $\widehat{\theta} = 0$.

Results

In both conditions with exposure control, the models for the shadow test completed with the ineligibility constraints always had a feasible solution. Thus, the independence assumption in (11) was automatically met for all items.

Figure 1 shows the exposure rates of the discrete items and stimuli in the pool for the three conditions after 10,000 examinees. For the condition without exposure, 38 of the items and stimuli had exposure rates that seriously violated the target rate of .25. However, both conditions with exposure control showed perfect exposure rates. Except for a few random violations (smaller than .01), all items had exposure rates below the target value of .25.

[Figure 1 about here]

Observe that the distributions of exposure rates for the two conditions with exposure control are indistinguishable. Application of the technique of fading thus had no impact whatsoever on the distribution of the exposure rates. Also, from this figure it is obvious that the control of the exposure rates of the 38 items and stimuli with a tendency to overexposure led to an increase in the exposure rates of several of the other items and stimuli. This result is to be expected because the average exposure rate in the pool is always equal to nI^{-1}, where n is the length of the test and I the size of the pool (van der Linden, 2002, proposition 1).

To find out if the exposure rates for the method in this paper stabilized quickly, we also checked the rates of the items and stimuli after 1,000 examinees. Figures 2 and 3 show the distributions of the differences between the actual rates after 1,000 and 10,000 examinees for the condition without and with probability updates with fading, respectively. For both conditions nearly all differences were smaller than 1%, indicating that the method was already stable after 1,000 examinees.

[Figures 2 and 3 about here]

The bias and MSE functions for the ability estimator estimated from 10,000 simulated examinees are given in Figures 4 and 5, respectively. The differences between these two functions for the three conditions are negligible for all practical purposes. The method of item exposure control in this paper thus had negligible impact on the statistical quality of the ability estimator.

[Figures 4 and 5 about here]

From these results it seems safe to conclude that the method works well. The distribution of the actual exposure rates was exactly as desired and already stable after 1,000 examinees. The fact that the impact of the methods on the statistical properties of the ability estimator was negligible is easily explained by the model for the assembly of the shadow tests in (1)-(2). The MSE of the ability estimator depends directly on the value of the objective function for the optimal shadow test. This value, in turn, depends on the number and severeness of the constraints. For real-life CAT programs, the number of ineligibility constraints that have to be added to the model is expected to be small relative to the number of content constraints. For example, for the LSAT the number of content constraints was 433, whereas we only had to constrain the exposure rates of 38 items to a value below .25. The average exposure rate of these items in the condition without exposure control was .70, which implies that on average for each examinee some 17 ineligibility constraints had to be added to the model, which implied an increase in the number of constraints by only 3.9 percent. If the item pool is well designed, ineligibility constraints are not severe and their impact on the value of the objective function for the solution is negligible.

Discussion

A practical advantage of the method of exposure control introduced in this paper is that, unlike the Sympson-Hetter method, it does not need a long iterative process of setting values for control parameters. The adaptive nature of the method guarantees that the

exposure rates are automatically set, and the method can be applied directly to operational tests.

The adaptive nature of the method also opens up a whole new range strategies of item pool management. For example, if one or more items are detected to be flawed or have been the victim of security breaches, these items can simply be removed from the pool or replaced by other items, and the exposure rates for the items in the new pool are adapt automatically to below the target value. For the Sympson-Hetter method, the pool would have to be taken out of operation to set new values for their control parameters (Chang & Harris, 2002).

We could also use such replacements intentionally, and periodically replace parts of the item pool for which the absolute numbers of exposures have reached a predetermined level. This strategy is an alternative to item-exposure control strategies based on random rotation of multiple item pools (Mills & Steffen, 2000). As a matter of fact, it can be seen as a rotation method that permanently rotates different item pools among the examinees. The rotation scheme does not require any physical transport of items. Neither does it require the assembly of multiple item pools with prior estimation of required overlap rates (Stocking & Swanson, 1998). Also, the size of the rotating pools as well as their item overlap rates are automatically set at optimal levels.

As a final example of a new strategy of item pool management, observe that there is no need to set the maximum exposure rate r_{max} at a common values for all items and examinees. This target could be set at different values for different set of items, for example, at lower values for items or stimuli that are easier to memorize or for item pools that are used at locations with a higher risk of item compromise. Also, the target could be set at different levels during the history of an individual item pool. These and other strategies for optimizing item pool management deserve further investigation.

As already noted, the method presented in this paper can also be applied conditional on a series of ability levels. Formally, the only thing required to get constraints on conditional item-exposure rates is to formulate all probabilities conditional on θ. Practically, in a conditional version of the method, the estimates of the conditional

probabilities of item eligibility in (19) are still updated after the examinee has completed the test, but the eligibility experiments with these probabilities are conducted more than once for each examinee, namely at each θ value at which the item-exposure rates are to be controlled. Currently, a study is conducted in which, analogous to the one for the Sympson-Hetter method in Stocking and Lewis (2000), the possible impact of estimation error in the ability estimates during the test on the actual conditional item-exposure rates is assessed.

References

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive tests.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Eggen, T. J. H. M. (2001). *Overexposure and underexposure of items in computerized adaptive testing* (Measurement and Research Department Reports 2001-1). Arnhem, The Netherlands: Citogroep.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

ILOG, Inc. (2000). *CPLEX 6.6* [Computer program and manual]. Incline Village, NV: Author.

Kingsbury, G. G., & Zara, A. R. (1991). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*, 359-375.

Jensen, F. V. (2001). *Bayesian networks and graphs*. New York: Springer.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 223-226). New York: Academic Press.

Mills, G. N., & Steffen, M. The GRE adaptive test: Operational issues. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 75-99). Boston: Kluwer.

Parshall, C., Hogarty, K., & Kromrey, J. (1999, June). *Item exposure in adaptive tests: An empirical investigation of control strategies*. Paper presented at the Annual Meeting of the Psychometric Society, Lawrence, KS.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item-exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 311-327.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics,*

23, 57-75.

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163-182). Boston: Kluwer.

Stocking, M. L., & Swanson, D. Optimal design of item banks for computerized adaptive testing. *Applied Psychological Measurement, 22*, 271-279.

Swanson, D., & Stocking, M. L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277-292.

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thomasson, G. L. (1995, June). *New item exposure control algorithms for computerized adaptive testing*. Paper presented at the Annual Meeting of the Psychometric Society, Minneapolis, MN.

van der Linden, W.J. (Ed.) (1998). Optimal test assembly of psychological and educational tests. *Applied Psychological Measurement, 22*, 195-211.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27-52). Boston: Kluwer.

van der Linden, W. J. (2002). *A formal characterization and some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing*. Submitted for publication.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1-25). Boston: Kluwer.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259-270.

van der Linden, W. J., & Veldkamp, B. P. (2002). *Implementing Sympson-Hetter item-*

exposure control in adaptive testing with shadow tests. Manuscript to be submitted for publication.

van der Linden, W.J., Veldkamp, B.P., & Reese, L.M. (2000). An integer programming approach to item pool design. *Applied Psychological Measurement, 24*, 139-150.

Veldkamp, B. P. (2001). *Implementing constrained CAT with shadow tests for large item pools*. Submitted for publication.

Veldkamp, B.P., & van der Linden, W.J. (2000). Designing item pools for computerized adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 149-162). Boston, MA: Kluwer Academic Publishers.

Figure Captions

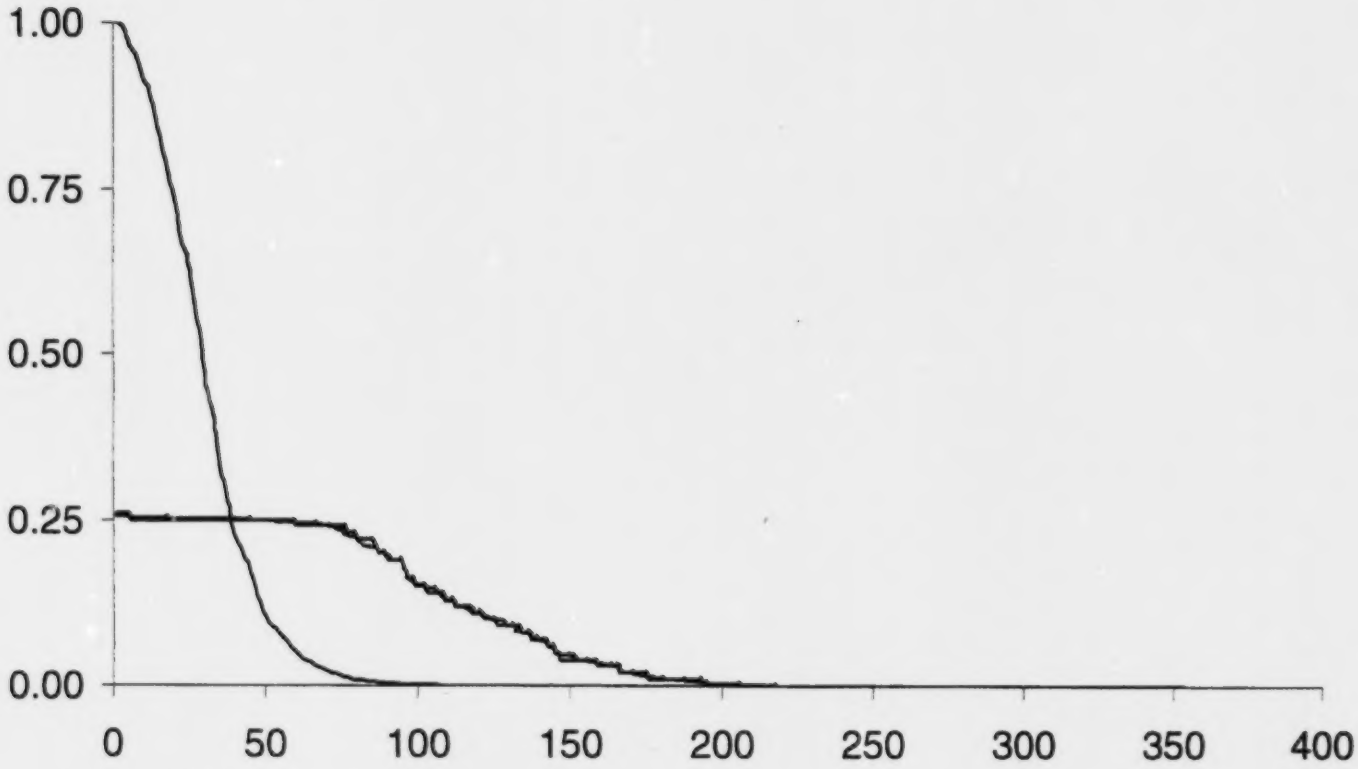
Figure 1. Distribution of exposure rates for the items and stimuli in the pool for CAT without exposure control (solid line), with exposure control (dashed line) and with exposure control and probability updates based on fading (dotted line).

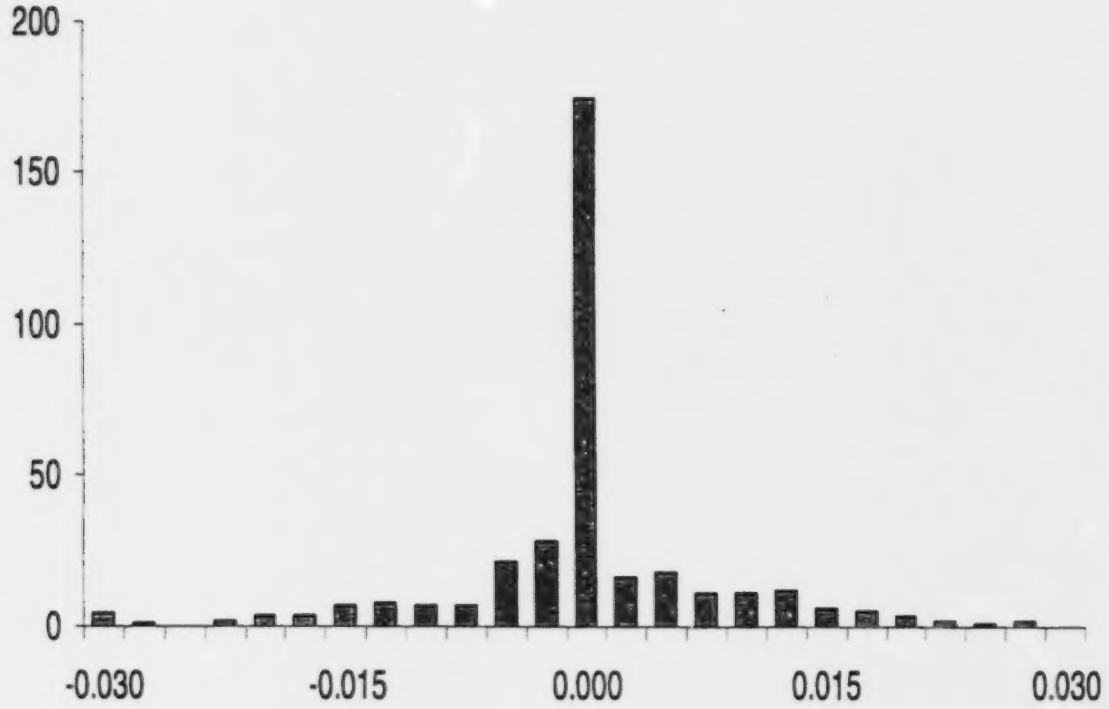
Figure 2. Differences between exposure rates after 1,000 and 10,000 examinees for the CAT with exposure control.

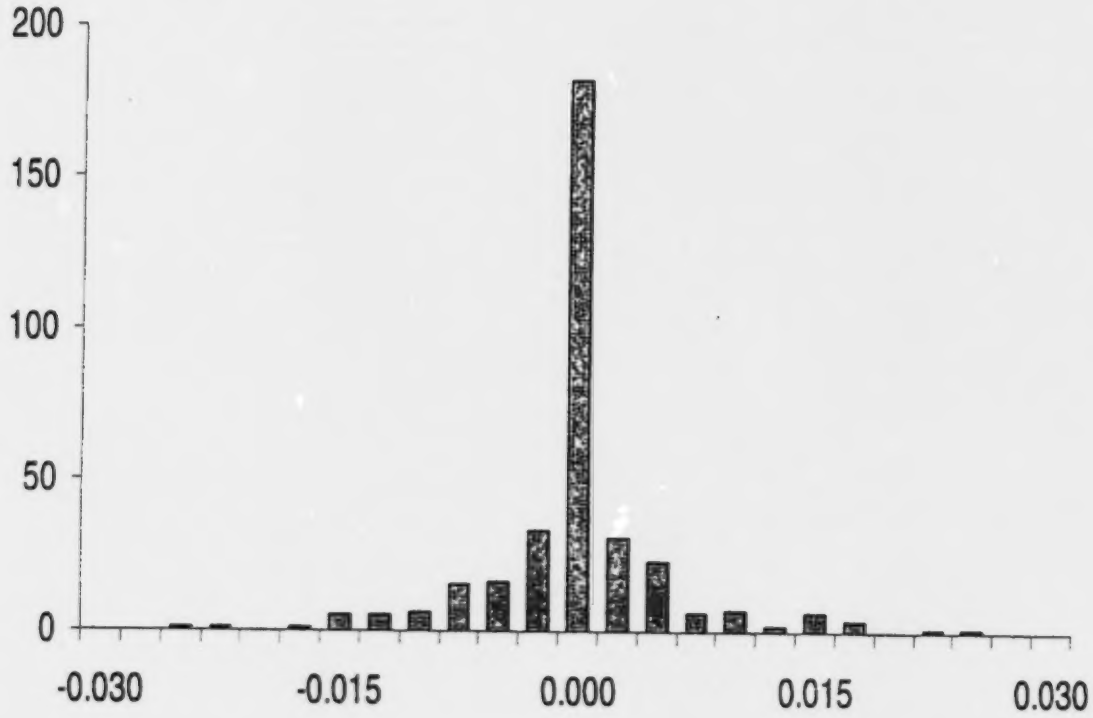
Figure 3. Differences between exposure rates after 1,000 and 10,000 examinees for the CAT with exposure control and probability updates based on fading.

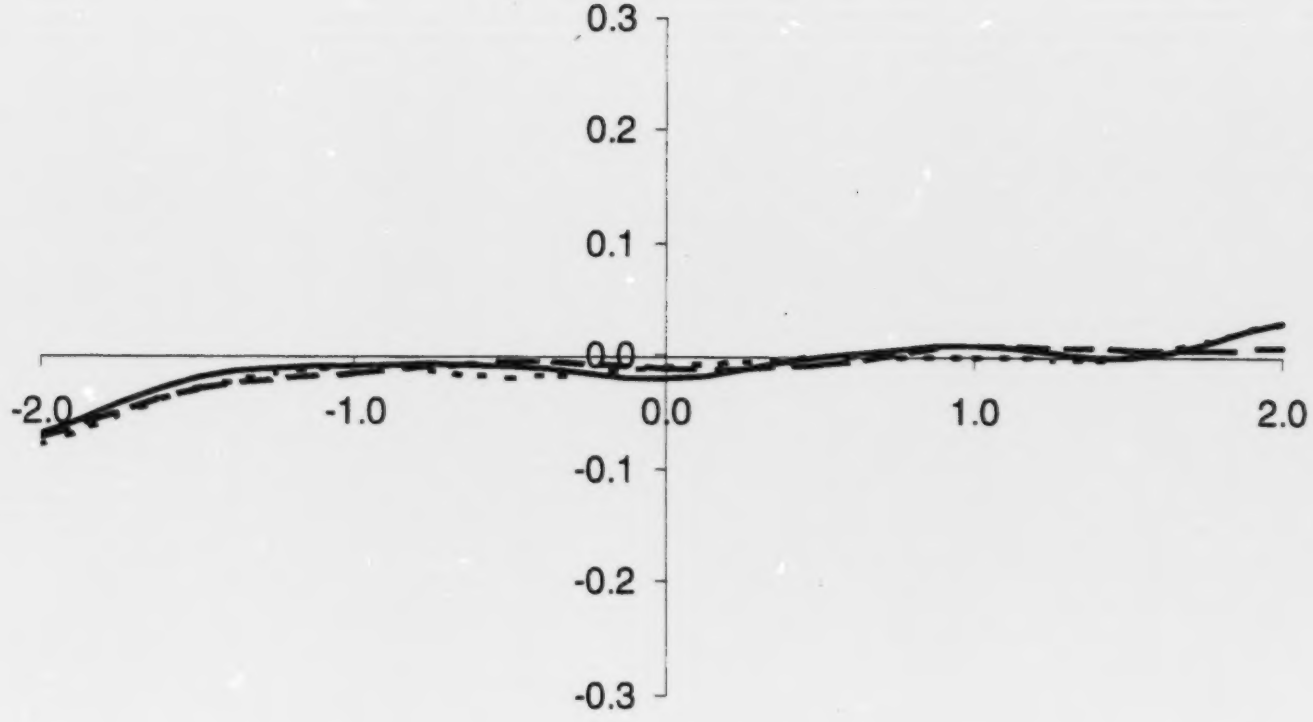
Figure 4. Estimated bias functions of the ability estimator for CAT without exposure control (solid line), with exposure control (dashed line) and with exposure control and probability updates based on fading (dotted line).

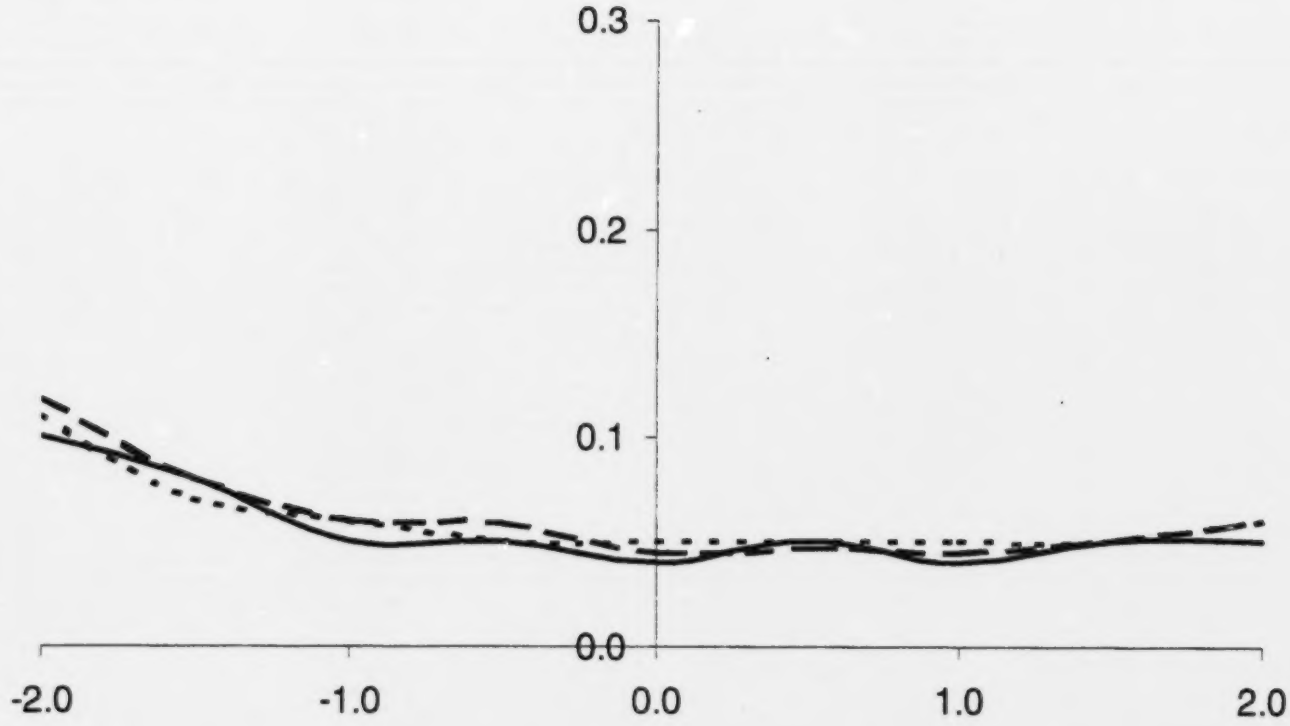
Figure 5. Estimated MSE functions of the ability estimator for CAT without exposure control (solid line), with exposure control (dashed line) and with exposure control and probability updates based on fading (dotted line).


1.00
0.75
0.50
0.25
0.00
0
50
100
150
200
250
300
350
400


200
150
100
50
0
-0.030
-0.015
0.000
0.015
0.030


200
150
100
50
0
-0.030
-0.015
0.000
0.015
0.030


0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-2.0
-1.0
0.0
1.0
2.0


0.3
0.2
0.1
0.0
-2.0
-1.0
0.0
1.0
2.0

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-02-06 W.J. van der Linden & B.P. Veldkamp, *Constraining Item Exposure in Computerized Adaptive Testing with Shadow Tests*

RR-02-05 A. Ariel, B.P. Veldkamp & W.J. van der Linden, *Constructing Rotating Item Pools for Constrained Adaptive Testing*

RR-02-04 W.J. van der Linden & L.S. Sotaridona, *A Statistical Test for Detecting Answer Copying on Multiple-Choice Tests*

RR-02-03 W.J. van der Linden, *Estimating Equating Error in Observed-Score Equating*

RR-02-02 W.J. van der Linden, *Some Alternatives to Sympson-Hetter Item-Exposure Control in Computerized Adaptive Testing*

RR-02-01 W.J. van der Linden, H.J. Vos, & L. Chang, *Detecting Intrajudge Inconsistency in Standard Setting using Test Items with a Selected-Response Format*

RR-01-11 C.A.W. Glas & W.J. van der Linden, *Modeling Variability in Item Parameters in Item Response Models*

RR-01-10 C.A.W. Glas & W.J. van der Linden, *Computerized Adaptive Testing with Item Clones*

RR-01-09 C.A.W. Glas & R.R. Meijer, *A Bayesian Approach to Person Fit Analysis in Item Response Theory Models*

RR-01-08 W.J. van der Linden, *Computerized Test Construction*

RR-01-07 R.R. Meijer & L.S. Sotaridona, *Two New Statistics to Detect Answer Copying*

RR-01-06 R.R. Meijer & L.S. Sotaridona, *Statistical Properties of the K-index for Detecting Answer Copying*

RR-01-05 C.A.W. Glas, I. Hendrawan & R.R. Meijer, *The Effect of Person Misfit on Classification Decisions*

RR-01-04 R. Ben-Yashar, S. Nitzan & H.J. Vos, *Optimal Cutoff Points in Single and Multiple Tests for Psychological and Educational Decision Making*

RR-01-03 R.R. Meijer, *Outlier Detection in High-Stakes Certification Testing*

RR-01-02 R.R. Meijer, *Diagnosing Item Score Patterns using IRT Based Person-Fit Statistics*

RR-01-01 H. Chang & W.J. van der Linden, *Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test Approach*

RR-00-11 B.P. Veldkamp & W.J. van der Linden, *Multidimensional Adaptive Testing with Constraints on Test Content*

RR-00-10 W.J. van der Linden, *A Test-Theoretic Approach to Observed-Score Equating*

RR-00-09 W.J. van der Linden & E.M.L.A. van Krimpen-Stoop, *Using Response Times to Detect Aberrant Responses in Computerized Adaptive Testing*

RR-00-08 L. Chang & W.J. van der Linden & H.J. Vos, *A New Test-Centered Standard-Setting Method Based on Interdependent Evaluation of Item Alternatives*

RR-00-07 W.J. van der linden, *Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing*

RR-00-06 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory*

RR-00-05 B.P. Veldkamp, *Modifications of the Branch-and-Bound Algorithm for Application in Constrained Adaptive Testing*

RR-00-04 B.P. Veldkamp, *Constrained Multidimensional Test Assembly*

RR-00-03 J.P. Fox & C.A.W. Glas, *Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory*

RR-00-02 J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*

RR-00-01 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*

RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*

RR-99-07 N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*

RR-99-06 G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*

RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*

RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.

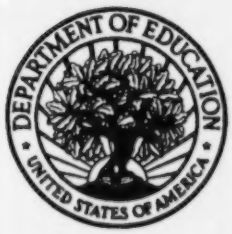
faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

COPY AVAILABLE

TM034741



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").